RESOLUTION AND RELIABILITY IN FUNCTIONAL CONNECTIVITY ANALYSIS

by
Jingran Zhu

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
April 2021

Abstract

Increasingly, researchers are interested in the functional connectivity between different brain regions of resting-state imaging. However, quantifying the reliability of measures of both brain function and structure is difficult, while reliability is essential to accurately detect differences between subjects or groups and predict outcomes. Here, we seek to evaluate the reliability and prediction performance of functional connectivity networks obtained using independent components analysis (ICA) with varying number of nodes. Reliability is measured by intra-class correlation coefficient (ICC), image intra-class correlation coefficient (I2C2) and Kolmogorov–Smirnov (KS) test. In particular, we evaluated how the number of components influence the results. We found the local (ICC) and global (I2C2) reliability have different trends with the change in the resolution of the ICA-based pacellation, and higher reliability was achieved on sessions that took place on the same day compared to different days for both ICC and I2C2. Individual patterns of functional connectivity do exist. First, KS test shown greater similarity between different sessions than different subjects. Second, fingerprinting and prediction of sex using connectivity matrix yield high prediction accuracy (over 90%). In addition, a moderate resolution of the parcellation seems to be an optimal choice. It provides more information about functional connectivity, improves the performance of prediction and avoids unnecessary computational cost, which only contribute little improvement to prediction accuracy.

Primary Reader and Advisor: Martin Lindquist

Secondary Reader: Brian Caffo

Acknowledgements

I wish to express my sincere appreciation to my academic advisor Professor Martin Lindquist. He has helped me with great patience and extensive knowledge in the past several months. He has always been willing to listen to my ideas and leave me much independent space to realize my ideas in research. In addition, he has motivated me and offered helpful advice when I faced with some difficulties. I learned a lot of statistical knowledge in the field of imaging from him. Without his help, my thesis would not have such abundant contents as it has.

I wish to thank Professor Brian Caffo who is the second reader of this thesis, for his detailed and valuable comments. I am grateful for the chance to work with and learn from excellent researcher as him.

I wish to thank John Muschelli, Elizabeth Colantuoni and my family, who gave me consistent help in the past two years. Without their support, I would not have achieved so much.

Contents

List of Tables

List of Figures

1 Introduction

Reliability is the ability to obtain similar findings in repeated measurements of a particular metric of interest. Quantifying the reliability of brain imaging measures is difficult, due to the complex nature of the data. However, increasingly researchers are seeking to determine the reliability of measures of both brain function and structure [3]. This is important, as a lack of reliability can hinder the ability to use these measures to accurately detect differences between subjects or groups and predict behavioral or clinical outcomes [10].

In the context of the study of brain function, researchers often use measures derived from functional magnetic resonance imaging (fMRI). Here one can either focus on measures of task-based activation or edge-level functional connectivity networks. There are various methods to assess the reliability. When working with task-based activation maps, some use a threshold to assess thhe reliability of significant voxels only, while others take similarity over all voxels into considerations [9]. Importantly, methods that leverage a threshold, such as Dice coefficient and Jaccard index [30], are highly dependent on the threshold that used to define the significant voxels. When assessing the reliability of untresholded activation maps, the most standard method used for assessing the stability of the results in replications of scans is the intraclass correlation (ICC). This measure is computed separately for each location in the brain which can make it difficult to conclude a global reliability for the whole image. In contrast, the Image Intra-Class Correlation (I2C2) coefficient generalizes the classic ICC to the case when the data of interest are images, or alternately entire networks, and therefore serves as a global measure of reliability for imaging studies [35].

Increasingly, researchers are interested in resting-state functional connectivity, or the

undirected association between different brain regions, measured while the subject is at rest. Here the brain is split into a number of regions of interest (ROIs), and the correlation between all pairwise combinations of ROIs are assessed. In this context, the ROIs are typically referred to as the nodes, and the correlation between them as the edges of the corresponding network. When assessing the reliability of functional connectivity networks, one can either compute ICC for each edge or I2C2 for the entire network.

A related topic to reliability, is so-called functional connectivity fingerprinting [13]. Here one assesses the ability to identify subjects from a large group based solely on their functional connectivity profiles. The success of this approach has led to a growing body of research showing that a person's functional connectivity act as a brain "fingerprint", indicating a measure of its reliability. Similarly, a rank sum styled test called "discriminability" is used to define reliability [26]. However, in both task and rest-state fMRI, systematic differences exist in group-mean connectivity, and high variability is present in each individual's functional connectivity profile [16, 27]. Researchers have also found that the reliability of functional connectivity varies across different networks in the brain.

At the same time, significant individual differences also exist in the connectivity, reflecting intrinsic differences in the brain activity and structure [33]. These individual differences offer hope that one can use functional connectivity profiles to predict behavioral variables or clinical outcomes. For example, [11] used functional connectivity to classify subjects according to sex. Sex differences in functional connectivity are important as there are observed sex differences in brain disorders and structure. For example, functional connectivity is stronger for males in unimodal sensorimotor cortices, and for females in the default mode network [29]. Constructing a classifica-

tion model for sex using functional connectivity could contribute to the understanding of the between-sex variation in brain activities.

When computing functional connectivity, an important choice is how to split, or parcellate, the brain into nodes. Parcellations can be based on anatomical or functional brain properties, and they can either be pre-defined or computed in a data-driven manner [38]. It is common for researchers to use pre-defined brain atlases that define a set of ROIs that cover the whole brain [24, 41, 34]. However, different atlases are not mutually consistent [5], and they may not all fit the data well. Ultimately, ROI-based analysis depends heavily on the choice of ROIs, and the wrong choice can adversely affect downstream analysis. In contrast, data-driven brain parcellations often perform better. These parcellations are typically derived using rs-fMRI data [43, 4, 7]. Here we focus on data-driven functional parcellations such as independent components analysis (ICA) and variants of principal components analysis (PCA) [21, 42, 1]. These parcellations tend to focus on determining components made up of distributed networks of regions, rather than individual regions in isolation. An important choice is deciding on the number of components to compute, which is directly related to the granularity of the determined nodes. Therefore, in this work we seek to evaluate the reliability and prediction performance of functional connectivity networks obtained using ICA as a data-driven approach towards defining nodes. In particular, we will evaluate how the number of independent components (nodes), that are defined influence the results. This is related to the granularity of the nodes and could serve as guidelines for choosing the size of the regions used in brain parcellations.

We use 6 different ICA-based parcelations computed at different resolutions ranging from 15 regions to 300 regions. For each resolution, we use test re-test resting-state fMRI data to access the reliability of the computed functional connectivity networks

using a variety of metrics. In addition, we access how well an individual's function connectivity networks is able to predict their sex. We seek to explore these different questions to determine the optimal resolution to use when performing ICA-based parcellation of the brain.

2 Methods

2.1 Data

We used the preprocessed and artifact-removed rs-fMRI data as provided by the 900 subject data release. The preprocessing and the artifact-removing procedures performed are explained in detail elsewhere [15, 37, 17, 31], and briefly described below. Each run was minimally preprocessed [15, 37], and artifacts were removed using the Oxford Center for Functional MRI of the Brain's (FMRIB) ICA-based X-noiseifier (ICA + FIX) procedure [17, 31]. At this point in the processing pipeline, rs-fMRI data from each run were represented as a time series of grayordinates, a combination of cortical surface vertices and subcortical standard-space voxels [15]. Each run was temporally demeaned and variance normalized [2]. All four runs for 461 subjects were fed into MELODIC's Incremental Group-Principal Component Analysis (MIGP) algorithm, which estimated the top 4500 weighted spatial eigenvectors. GICA was applied to the output of MIGP using FSL's MELODIC tool [2] using six different dimensions (i.e., number of independent components: 15, 25, 50, 100, 200, 300).

Dual-regression was then used to map group-level spatial maps of the components onto each subject's time series data [12]. For dual-regression, the time series of each of the runs were first concatenated within subjects in the following order: Day 1 LR, Day 1 RL, Day 2 LR, Day 2 RL. Then the full set of group-level maps were used as spatial regressors against each subject's full time series (4800 volumes) to obtain a single representative time series per IC. The functional assignment of each component was determined as described above (refer to section 2.2.1) using the Allen RSNs. Subject- and run-specific time series from the components then served as input for our analyses. For each dimension d we computed a $d \times d$ correlation matrix consisting of all pairwise correlations between nodes

As shown in **Table 1**, we have 820 subjects in total for this project, consisting of 453 (55%) females and 367 (45%) males. The mean age is 28.8, mean height is 67.3 inches, mean weight is 171 lb and mean BMI is 26.4 $kg\backslash m^2$. People with a BMI of 26 to 27 is about 20 percent overweight, which is generally believed to carry moderate health risks.

Table 1
Subject demographics by sex, including age, height, weight and BMI

	Female (N=453)	**Male (N=367)**	**Total (N=820)**
Age			
Mean (SD)	29.4 (3.59)	28.0 (3.72)	28.8 (3.71)
Median [Min, Max]	30.0 [22.0, 36.0]	28.0 [22.0, 37.0]	29.0 [22.0, 37.0]
Height (inches)			
Mean (SD)	64.7 (3.99)	70.5 (2.95)	67.3 (4.57)
Median [Min, Max]	65.0 [0, 72.0]	71.0 [63.0, 80.0]	67.0 [0, 80.0]
Weight (lb)			
Mean (SD)	156 (36.6)	190 (35.8)	171 (40.0)
Median [Min, Max]	149 [0, 305]	186 [124, 302]	166 [0, 305]
BMI			
Mean (SD)	26.1 (5.82)	26.9 (4.50)	26.4 (5.28)
Median [Min, Max]	24.8 [0, 47.8]	26.2 [18.4, 41.0]	25.4 [0, 47.8]

2.2 Assessment of Reliability

In this section, we discuss three methods to assess the reliability of functional connectivity networks where the nodes are derived using ICA of varying dimensions. These include ICC, which focuses on local reliability (i.e., the reliability of each element of the correlation matrix, and I2C2, which is a global measure of reliability. In addition,

we evaluate the equality of the density of the elements in the ICC matrix using the Kolmogorov–Smirnov test. We describe each approach below.

2.2.1 Intra-class Correlation Coefficient (ICC)

Since there are many different forms of ICC, selection of a correct form is essential. There are three potential models for ICC, one-way random-effects model, two-way random-effects model, and two-way mixed-effects model [22]. Since we assume our raters (scan sessions) are randomly selected from a larger population of raters with similar characteristics, the two-way random-effects model is used. As a result, we can generalize the reliability results to any scans which possess the same characteristics as those selected in the current study. For the type selection, since the measurement from a single rater is the basis of the actual measurement,"single type" is selected instead of "mean". For the ICC definition, "absolute agreement" is selected, which is related to whether different sessions assign the same score to the same subject, while "consistency" describes whether different sessions' scores to the same group of subjects are correlated [25].

Let X_i be the true image, which is assumed to be independent across subjects. The W_{ij} are the proxy measurements of the X_i. The random variables U_{ij} represent measurement error, which are assumed to be independent across subjects and replicates and are independent of X_i. The classical measurement error model [6, 14] in replication studies is

$$W_{ij} = X_i + U_{ij}, \tag{1}$$

where for i stands for subject (i= 1, ..., 820) and j stands for replication (session) (j=1, ..., 4). We assumed the variance of X_i is σ_X^2, and U_{ij} have the same variance of σ_U^2. For the subject i, W_{ij} are correlated, its correlation is the intra-class correlation

coefficient (ICC) , which equals to

$$\text{corr}\,(W_{i1}, W_{i2}) = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} = \frac{\sigma_W^2 - \sigma_U^2}{\sigma_W^2} = 1 - \frac{\sigma_U^2}{\sigma_W^2}.$$

Using ICC, one could measure the reliability of each component of the correlation matrix for the same subject. Of note, here the W_{ij} are transformed versions of the observed components of the correlation matrix, where the Fisher r-to-z transformation is applied to approximate normality for the upper diagonal entries of the correlation matrix.

Given the image is divided into R regions (R=15, 25, 50, 100, 200, 300), its correlation is a R by R matrix and the upper triangle of the correlation matrix without the diagonal has $\frac{R\times(R-1)}{2}$ components. Each region of the image belongs to ten different brain networks, which are Visual, Somatomotor, Dorsal Attention, Ventral Attention, Limbic, Frontoparietal, Default Mode Network (DMN), Basal Ganglia, Cerebellum and Brainstem. The ICC results of all components are placed at their original position on the upper triangle of correlation matrix, which are then ordered by regions' belonging networks. In addition, symmetric full-matrix heat map is generated based on the ICC in the form of upper triangle so that characteristics of the reliability of the correlations within and between the network can be observed more easily.

2.2.2 Image Intra-class Correlation Coefficient (I2C2)

Let $X_i(v)$ be the true image and $W_{ij}(v)$ be the proxy measurements of $X_i(v)$ at voxel v. All images are represented as $V \times 1$ vectors. Then the image measurement error is

$$W_{ij}(v) = X_i(v) + U_{ij}(v), \tag{2}$$

where i stands for subject ($i = 1, \ldots, 820$), and j stands for replication (session) ($j = 1, \ldots, 4$). The $\boldsymbol{W}_{ij} = \{W_{ij}(v) : v = 1, \ldots, V\}$ are the vectors of proxy images (which are observed), while $\boldsymbol{X}_i = \{X_i(v) : v = 1, \ldots, V\}$ are the true images (which are unknown). The $\boldsymbol{U}_{ij} = \{U_{ij}(v) : v = 1, \ldots, V\}$ are measurement error, assumed to be independent across subjects. K_W is the covariance of $\boldsymbol{W}_{ij}$, K_X is the covariance of $\boldsymbol{X}_i$, and K_U is the covariance of $\boldsymbol{U}_{ij}$, that is

$$\operatorname{cov}(\boldsymbol{W}_{ij}, \boldsymbol{W}_{ij}) = K_W,$$

$$\operatorname{cov}(\boldsymbol{X}_i, \boldsymbol{X}_i) = K_X,$$

and

$$\operatorname{cov}(\boldsymbol{U}_{ij}, \boldsymbol{U}_{ij}) = K_U,$$

where K_X is the within-subject covariance, and K_U is the covarience of the measurement error. Applying the application of multivariate variance operator to (1), we obtain $K_W = K_X + K_U$. Then the image intra-class correlation (I2C2) coefficient is defined as

$$\rho = \frac{\operatorname{trace}(K_X)}{\operatorname{trace}(K_W)} = \frac{\operatorname{trace}(K_W) - \operatorname{trace}(K_U)}{\operatorname{trace}(K_W)} = 1 - \frac{\operatorname{trace}(K_U)}{\operatorname{trace}(K_W)}. \tag{3}$$

To reduce the cost of computation, method of moments estimators is used to get

$$\widehat{\operatorname{trace}}(K_W) = \frac{1}{\sum_{i=1}^{I} J_i - 1} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{v=1}^{V} \left\{ W_{ij}(v) - \overline{\mathrm{W}}_{\cdot\cdot}(v) \right\}^2,$$

and

$$\widehat{\operatorname{trace}}(K_U) = \frac{1}{\sum_{i=1}^{I} (J_i - 1)} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{v=1}^{V} \left\{ W_{ij}(v) - \overline{\mathrm{W}}_{i\cdot}(v) \right\}^2.$$

$\overline{\mathrm{W}}_{\cdot\cdot}(v) = \Sigma_{i,j,v} W_{ij}(v)/IJ$ is the average of all images over all subjects and sessions and $\overline{\mathrm{W}}_{i\cdot}(v) = \sum_{j=1}^{J_i} W_{ij}(v)/J_i$ is the average image for subject i over all sessions j. The computational burden of calculating these estimates is linear in V. As a result, one can generate an estimate of I2C2 quickly.

We used an R package to calculate the I2C2, available at `http://www.biostat.jhsph.edu/~ccrainic/software.html`. The input matrix of the I2C2 function is a N by P data matrix. Each row contains the observed correlation data from a particular session for a single subject, obtained from the upper triangular portion of the correlation matrix. Again, Fisher r-to-z transformations were applied to ensure the normality of correlation matrix. Each column contains correlation values for all subjects and sessions. Here, $N = S \times I$ and $P = \frac{R\times(R-1)}{2}$, where S is number of sessions included (S=2, 4), I is the number of subjects (I=820), R is the number of regions of each image (R=15, 25, 50, 100, 200, 300).

To access the reliability of all sessions, correlation information from 4 sessions ($S = 4$) over 2 days is gathered as target of interest. To access the reliability of two sessions, correlation information from 2 sessions ($S = 2$), either on the same day or on different days is gathered, resulting in six combinations.

2.2.3 Kolmogorov–Smirnov Test

To investigate similarity of the functional connectivity across different sessions and different subjects, we performed the Kolmogorov–Smirnov test (KS test), a nonparametric test of the equality of continuous densities, on the density of the elements of the connectivity matrix. The D statistic of KS test is the absolute max distance between the cumulative distribution functions (CDFs) of the two samples. The closer

this number is to 0 the more likely it is that the two samples were drawn from the same distribution. For a given subject, we first averaged the D statistic of KS tests on different sessions for the same subject, then averaged the D statistic on different subjects for the same session. After that, two average D statistic were compared to investigate the similarity of different sessions and different subjects.

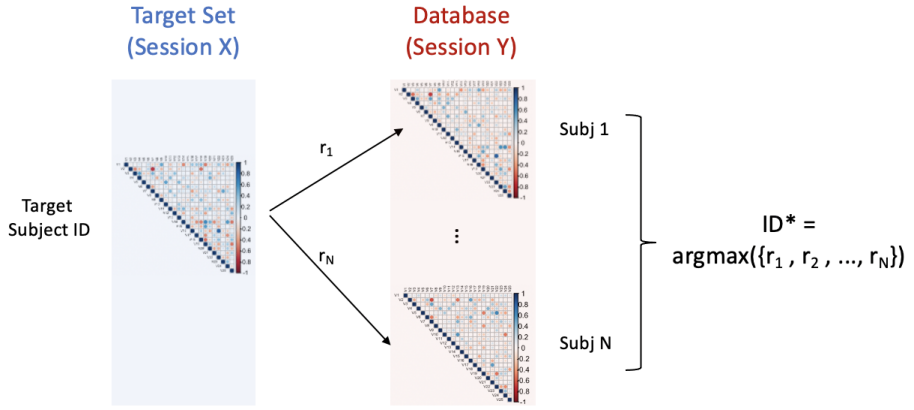
2.3 Functional Connectivity Fingerprinting

Next, we sought to investigate whether the functional connectivity patterns for each individual could serve as a "fingerprint" to help us identify that individual from a larger group. Identification was performed across pairs of scans, one as target and the other as database, obtained from different scan sessions. In total, wee have 4 sessions collected over 2 days.

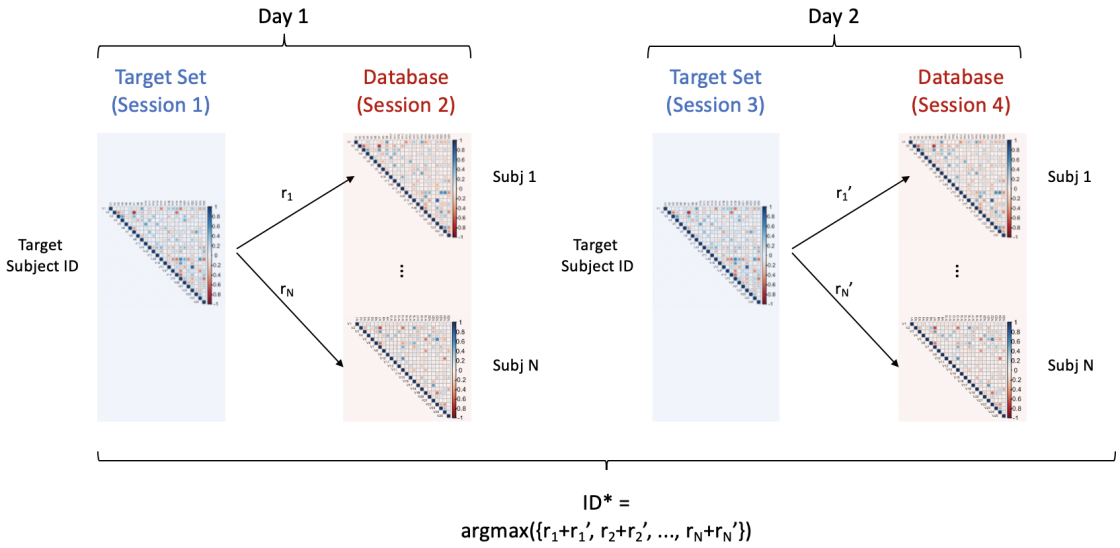
In an iterative process, we first chose the connectivity matrix of one individual (ID) from the target set and compared it with each of the connectivity matrices in the database to find the most similar one ($ID^* = \text{argmax}\left(\{r_1, r_2, \ldots, r_N\}\right)$), based on computing the Pearson correlation coefficient ($r_i, i = 1, 2, ..., N$) between the vectorized correlation matrices (**Fig.1(a)**). Next, we move on to the next individual in the target set and repeat the comparison above until we cover all individuals in the target set. Finally, we calculate the accuracy of identification based on how often we match to the same individual in the database set.

In addition, we applied another identification process by combining two sessions on the same day into one session. Sessions on one day serve as the target set and the sessions on the other day serve as the database set. With this combination, we used more information than the previous identification process. Similarly, we also combined two comparisons over two days as the identification rules (**Fig.1(b)**), which is the sum of

two Pearson correlation coefficients ($ID^* = \text{argmax}\left(\{r_1 + r_1', r_2 + r_2', \ldots, r_N + r_N'\}\right)$).



(a) Identification analysis by similarity of two sessions



(b) Identification analysis by sum of similarity of two sessions separately

Figure 1: (a) Identification analysis by similarity of two sessions. In an iterative process, we chose one individual's connectivity matrix from the target set as target subject with ID, which is compared with each in the database to find the most similar one ($ID^* = \text{argmax}\left(\{r_1, r_2, \ldots, r_N\}\right)$), which is decided by Pearson correlation coefficient between these two matrix ($r_i, i = 1, 2, ..., N$). (b) Identification analysis by sum of similarity of two sessions separately. Similarly, we chose target subject with ID from both session 1 and session 3, and compare them with each one in session 2 and session 4 separately to find the most similar one ($ID^* = \text{argmax}\left(\{r_1 + r_1', r_2 + r_2', \ldots, r_N + r_N'\}\right)$).

2.4 Prediction of Sex

Here we describe two techniques towards using the functional connectivity profiles to classify subjects according to their sex.

2.4.1 Regularized Logistic Regression Model

In this section, we applied logistic regularized regression models to predict a subject's sex based on their functional connectivity data, and compared the prediction accuracy with the full logistic regression model. With the help of regularized models, we could avoid the over-fitting problem, considering we will have $P = \frac{R\times(R-1)}{2}$ features in the functional connectivity data for image with R regions.

Lasso and ridge regression are two of most commonly used methods of regularized regression [19, 28, 32, 40]. Ridge logistic regression is obtained by maximizing the likelihood function, and a penalized parameter is applied to all coefficients except the intercept. Ordinary logistic regression is given by the probability of the response success

$$P\left(y_i = 1\right) = \pi_i = \frac{e^{x_\beta \beta}}{1 + e^{x_i \beta}}$$

where x_i is the i-th row of an matrix of n observations with p predictors and a column of ones to accommodate the intercept, y_i is the sex of subjects and β is the column vector of the regression coefficients. The ridge logistic regression estimator depends on the choice of a tuning parameter, $\lambda \geq 0$. The coefficients estimates are the values that maximize the log-likelihood function where a $L2$ ridge penalty is added [8, 23]:

$$l_\lambda^R(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \log\left(1 + e^{x\beta}\right)\right] - \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso logistic regression uses a $L1$ penalty instead of the $L2$ used in ridge regression. The $L1$ penalty used in the lasso is used for both variable selection and shrinkage,

since it can force some coefficient estimates to be equal to zero [20]. The lasso logistic regression could reduce the number of predictors in the final model, which in turn improves model interpretability. Its penalized version of the log-likelihood function to be maximized is [18]:

$$l_{\lambda}^{L}(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \log\left(1 + e^{x\beta}\right) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \,.$$

For functional connectivity data with R regions, prediction models are constructed within each scan session using lasso, ridge and full regression methods. First, 80% of the data is divided into a training set and the rest into a testing set. Prediction accuracy of sex using different models has been collected. Therefor, we evaluate the impact of regions, models and sessions on the prediction accuracy with functional connectivity information. The glmnet package in R was used to build the regularized models [36], and cross validation is used to select optimal λ in lasso and ridge regression models.

We are also interested in whether significance of β in lasso models has an association with the reliability of functional connectivity, which is calculated as ICC in previous section. As a result, symmetric full-matrix heat map is generated based on the absolute value of β, which is similar to the heat map of ICC in previous session, so that we can compare them directly. In addition, box plot is used for compare the average value and distribution of ICC for the features with significant β and non significant β.

2.4.2 Support Vector Machine

Support Vector Machines (SVMs), with roots in Statistical Learning Theory (SLT) and optimization methods, can be used as a tool for solving common problems in

machine learning, such as having finite training points and over-fitting [39]. Here we assume our functional connectivity data could not be fully separable by a linear boundary. In addition, we use a linear kernel to classify sex instead of nonlinear kernel to avoid over-fitting models and poor performance.

Given training vectors $\boldsymbol{x}_i \in R^n, i = 1, \ldots, l$, in two classes, and an indicator vector $\boldsymbol{y} \in R^l$ such that $y_i \in \{1, -1\}$, C-SVC (Boser et al., 1992; Cortes and Vapnik, 1995) solves the following primal optimization problem.

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i\left(\boldsymbol{w}^T\phi\left(\boldsymbol{x}_i\right) + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l, \end{aligned} \tag{4}$$

where $\phi\left(\boldsymbol{x}_i\right)$ maps $\boldsymbol{x}_i$ into a higher-dimensional space and $C > 0$ is the regularization parameter. For functional connectivity data with R regions, prediction models are constructed within each scan session using SVMs. First, 80% of data is divided into training set and the rest is the testing set. Prediction accuracy of sex using SVMs has been collected. The e1071 package in R was used to fit the SVMs.

3 Results

3.1 Reliability of Functional Connectivity Analysis

3.1.1 Intra-class Correlation Coefficient (ICC)

With ICC, we assess the reliability of each element in the functional connectivity matrix. In addition, each region belongs to a specific network in brain, and the different ICA-based parcellations consist of regions with differing compositions of the ten networks. As shown in **Fig.2**, regions from brain stem and cerebellum account for larger percentages, while regions from default mode network (DMN) and frontoparietal account for smaller percentages. We took networks into considerations so that we could refer to the functional connectivity reliability between specific networks.

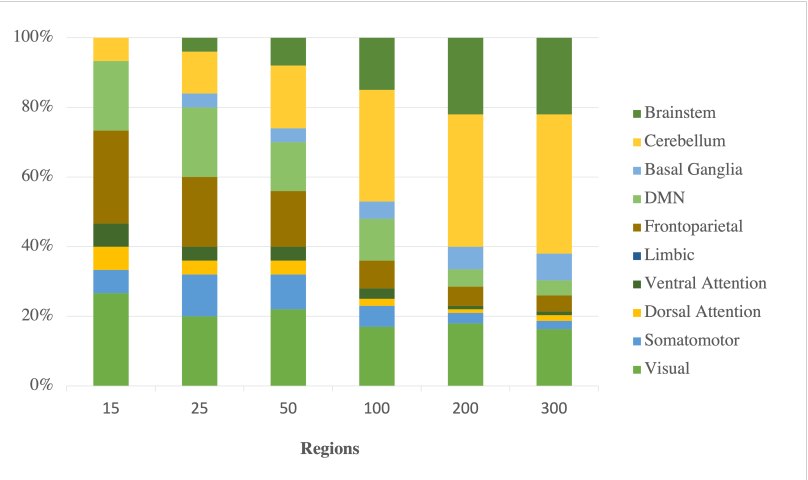
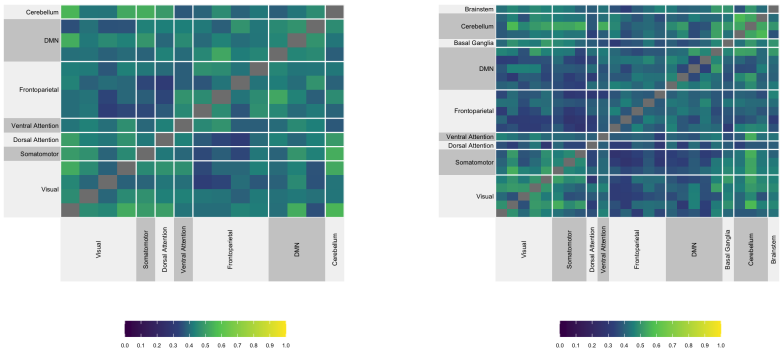


Figure 2: The composition of regions based on which network they belong to for each ICA-based parcellation. With the increase of regions from 15 to 300, the composition varies greatly. Regions from brain stem and cerebellum account for larger percentages, while regions from default mode network (DMN) and frontoparietal account for smaller percentages.

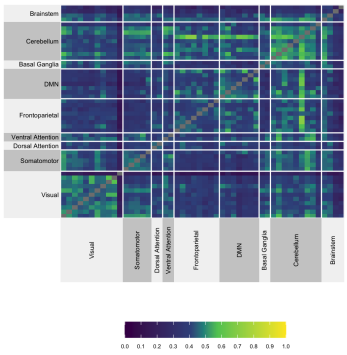
After computing the ICC for each component in the connectivity matrix, we ordered them by networks so that we could observe the patterns of reliability within and between networks (**Fig.3**). The patterns in the ICC matrices vary greatly with the change of dimesnions of the parcellation from 15 to 300 regions. For 15 regions, it

is hard to distinguish networks with high reliability of connectivity (yellow) between those with poor reliability (blue). For 50 and 100 regions, we are able to identify the networks with relatively high reliability, that is, their connectivity is consistent. A high ICC was observed within vision, cerebellum, frontoparietal and DMN. In addition, the connectivity between cerebellum and other networks has a relatively higher reliability. However, it becomes indistinguishable again when there are more than 200 regions, which may indicate that image is divided into more regions than needed.

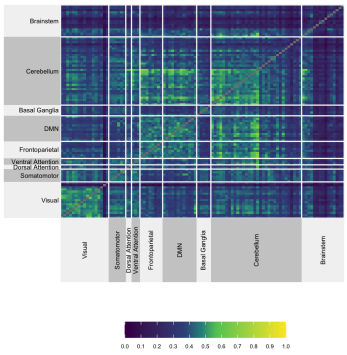


(a) $R = 15$

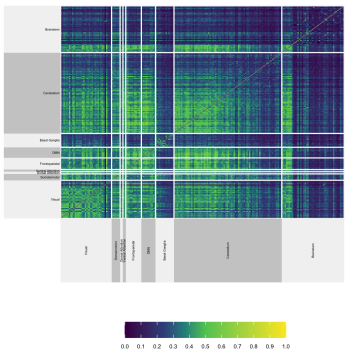
(b) $R = 25$



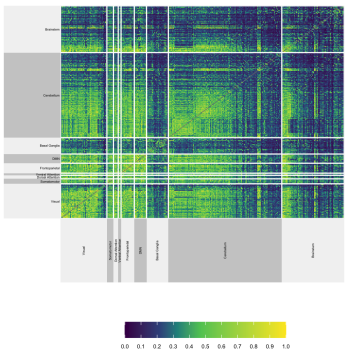
(c) $R = 50$



(d) $R = 100$



(e) $R = 200$



(f) $R = 300$

Figure 3: Heat map of ICC over all four sessions for varying resolutions. Blue stands for poor reliability of connectivity, while yellow stands for high reliability. When we have less regions, it is more homogeneous and is hard to distinguish the networks with high reliability between the ones with poor reliability. With more regions, we are able to identify the networks with relatively high reliability. High ICC was observed within vision, cerebellum, frontoparietal and DMN.

3.1.2 Image Intra-class Correlation Coefficient (I2C2)

Unlike ICC, I2C2 generates a global assessment of reliability for given regions and sessions instead of a matrix contains results for each component in the connectivity matrix. Considering we have four sessions over two days in total, we calculated I2C2 for seven combinations of sessions, which include either two sessions among four sessions or all four sessions. As shown in **Fig.4**, sessions on the same day (i.e., session 1 & session 2) yield relatively higher I2C2 than sessions on different days (i.e., session 1 & session 3). The highest I2C2 of over 0.5 is reached between session 3 and session 4 using the parcellation with 15 regions. However, the results of I2C2 are lower than expected, which may indicate the global reliability of functional connectivity between different scan sessions is quite low. Meanwhile, with the increase of regions from 15 to 300, I2C2 continuously decreases, which may be accounted by more variability induced. The relationship between regions and I2C2 is approximately linear for any combination of sessions (adjusted R-squared: 0.9883).

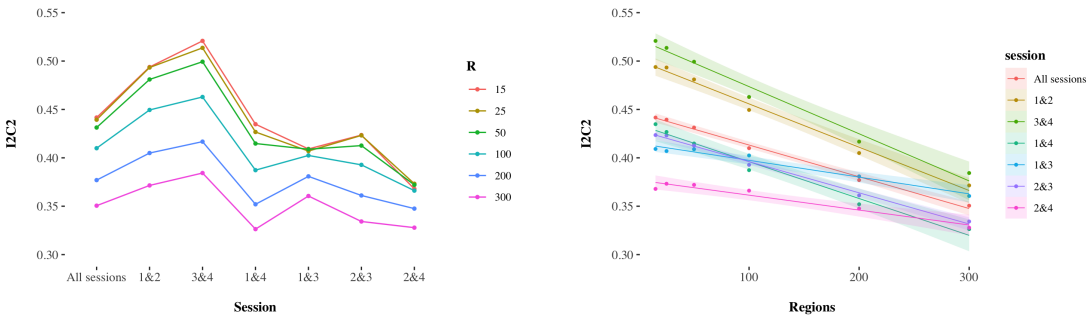


Figure 4: (a) I2C2 by sessions. I2C2 results for seven different combinations of sessions stratified by number of regions, R, including one combination of all four sessions over two days and six combinations of two sessions either on the same day (session 1&2, session 3&4) or on different days (session 1&4, session 1&3, session 2&3, session 2&4). Applying two sessions on the same day yields the highest I2C2. (b) I2C2 by resolution. I2C2 results of six different numbers of regions (R) used to divide the image, from 15 to 300 regions. Different colors stand for different sessions included in the calculation of I2C2. Linear smooth lines with standard error bounds were added. With the increase of number of regions, I2C2 continuously decreases for all kinds of combinations of sessions. The relationship between regions and I2C2 is approximately linear for any combination of sessions.

3.1.3 Kolmogorov–Smirnov Test

For functional connectivity similarity, average D statistic, the absolute max distance between the CDFs of the two samples, is greater when comparing different subjects than different sessions. It indicates that differences between subjects in the same session are greater than differences between different sessions for the same subject. This conclusion remains consistent when the number of regions varies from 15 to 300. With the increase of regions, D statistic becomes closer to 0, which means connectivity matrices are more similar between different sessions and subjects.

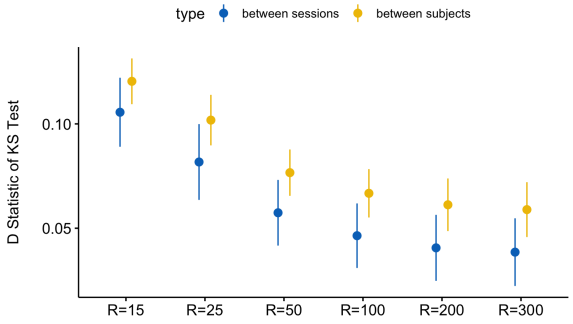


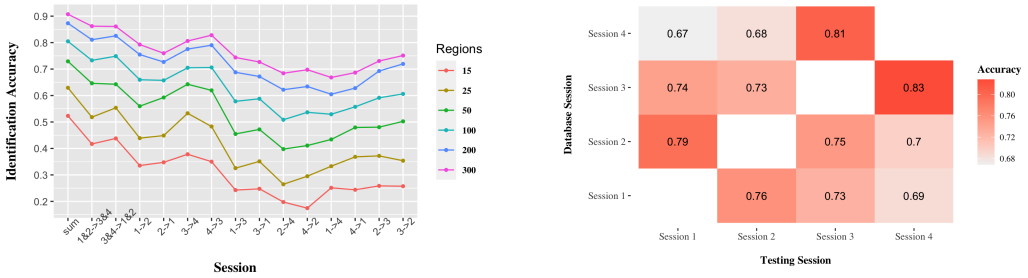
Figure 5: Mean D statistic of KS test from comparing the the connectivity matrix of different sessions and different subjects. D statistic is greater when comparing different subjects than different sessions. With the increase of regions, D statistic becomes closer to 0.

3.2 Functional Connectivity Fingerprint

As shown in **Fig.6(a)**, we applied fifteen different approaches towards identifying subjects. Switching the testing set and database set will yield slightly different identification accuracy, which means the identification process is not extremely symmetric. The method that combine two comparisons over two days as the identification rules (see **Fig.1(b)**) yields the highest identification accuracy of over 90% using the parcellation with 300 regions. And the method using two sessions on one day as the target set and sessions on the other day as the database set also gives relatively high identification accuracy. In terms of impact of regions, the accuracy continuously increases

with the increase in the number of regions.

In addition, the accuracy of methods using sessions on the same day is higher than the ones using sessions on different days, which is consistent with the reliability measurements observed using I2C2. This can also be observed in **Fig.6(b)**, showing that the accuracy is higher along the diagonal, which is consistent with that the combination of sessions along the diagonal were performed in the same day (i.e., session 1 & session 2, session 3 & session 4). And the identification accuracy is quite symmetric along the diagonal, which means that switching the testing and database session will only lead to small changes on identification performance.



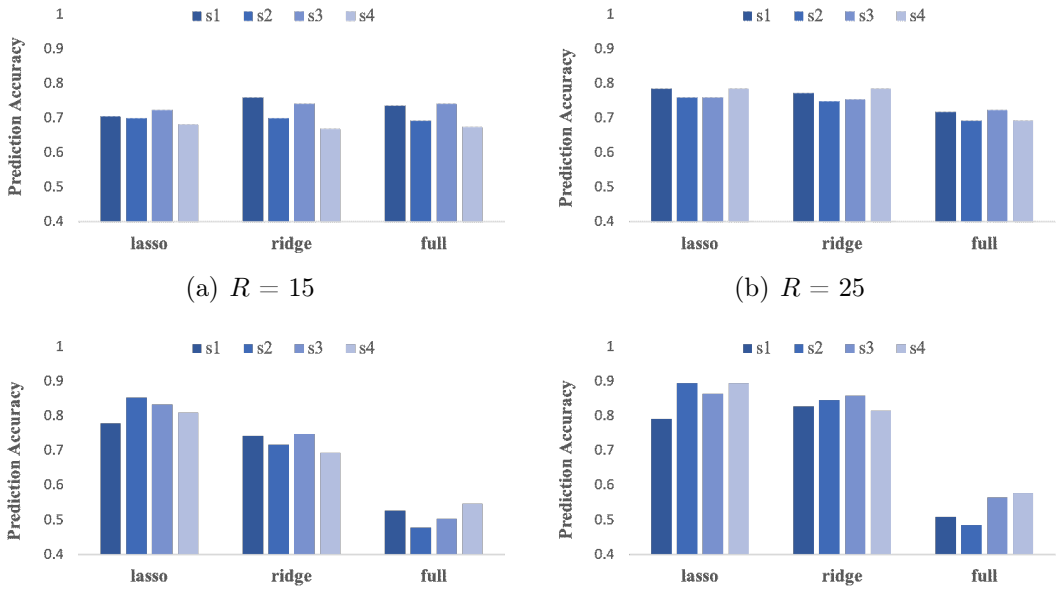
(a) (b)

Figure 6: (a) Identification accuracy colored by regions. The x-axis represents different choices of testing and database set. "sum" is the fingerprinting method illustrated in **Fig.1(b)**, which has the highest accuracy among all methods, while other methods are shown in **Fig.1(a)**. The accuracy of methods using sessions on the same day is higher than the ones using sessions on different days. With the increase in the number of regions, the accuracy continuously increases. (b) Heat map showing the identification accuracy with 300 regions. The x-axis is the testing session, while the y-axis is the database session. Higher accuracy is observed along the diagonal, consistent with that the combination of sessions along the diagonal were performed in the same day (i.e., session 1 & session 2, session 3 & session 4).

3.3 Prediction of Sex

Besides fingerprinting, we also constructed prediction models of sex using functional connectivity data. For 15 and 25 regions, two penalized regression models, lasso and

ridge, yield similar prediction accuracy with the full models (**Fig.7(a)** and **Fig.7(b)**). For given models and regions, prediction performance is quite consistent among the four sessions. For 50 and 100 regions, full models fails to predict sex accurately (**Fig.7(c)** and **Fig.7(d)**), since there are so many predictors that lead to overfitting problems with the increase of regions.



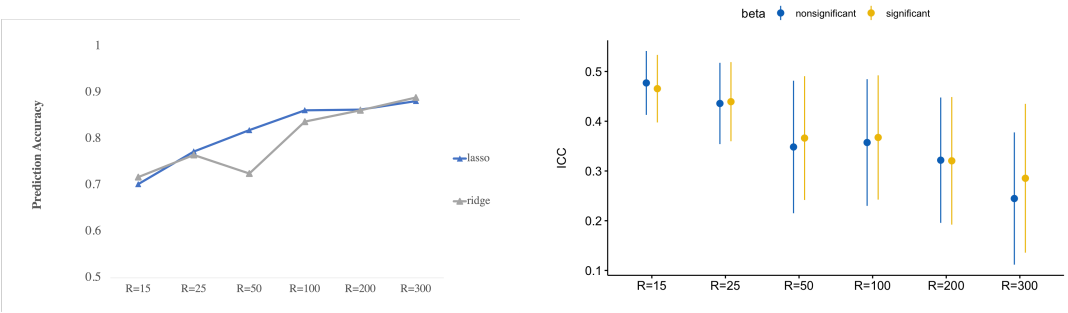
(a) $R = 15$ (b) $R = 25$

(c) $R = 50$ (d) $R = 100$

Figure 7: Prediction accuracy of sex with or without penalized regression models for different sessions using different regression models.. For 15 and 25 regions, two penalized regression models, lasso and ridge, yields similar prediction accuracy with the full models. For given models and regions, prediction performance is quite consistent among four sessions. For 50 and 100 regions, full models fails to predict sex accurately. Lasso and ridge models continuously improve their prediction performance.

On the other hand, lasso and ridge models continuously improve their prediction performance to over 90%. As shown in **Fig.8(a)**, mean Prediction accuracy of lasso model increased from 70% to 88%, while the mean prediction accuracy of ridge model increased from 72% to 89%. However, the increase after 100 regions only bring little improvement to prediction, which may indicate 100 regions are able to provide the majority of features for prediction of sex. For lasso models **Fig.8(b)**, with the region increase from 15 to 300, average ICC for features with both significant β and non

significant β decrease greatly. In addition, considering more features were included, the variance of ICC increase as expected. For 25, 50, 100 and 300 regions, it seems that features with significant β have slightly greater average ICC than the ones with non significant β. However, for 15 regions, features with significant β have slightly smaller average ICC than the ones with non significant β. As a result, it may indicate that more reliable features are not necessarily significant in prediction with lasso models.



(a) (b)

Figure 8: (a) Mean of prediction accuracy of sex using lasso and ridge model. For both penalized models, prediction accuracy increased with the increase of regions from 70% to near 90%. However, the increase after 100 regions only bring little improvement to prediction, which may indicate 100 regions are able to provide the majority of features for prediction of sex. (b) Box plot of ICC for features with significant β and non significant β. With the region increase from 15 to 300, average ICC for features with both significant β and non significant β decrease greatly. In addition, considering more features were included, the variance of ICC increase as expected.

Besides penalized regression models, we also used Support Vector Machines (SVMs) to predict sex. As shown in **Fig.9**, the prediction accuracy increase from 73% to 88% when regions increase from 15 to 100. Given certain regions, prediction accuracy is consistent among the four sessions. In addition, the prediction performance of SVM is consistent with those of penalized models.

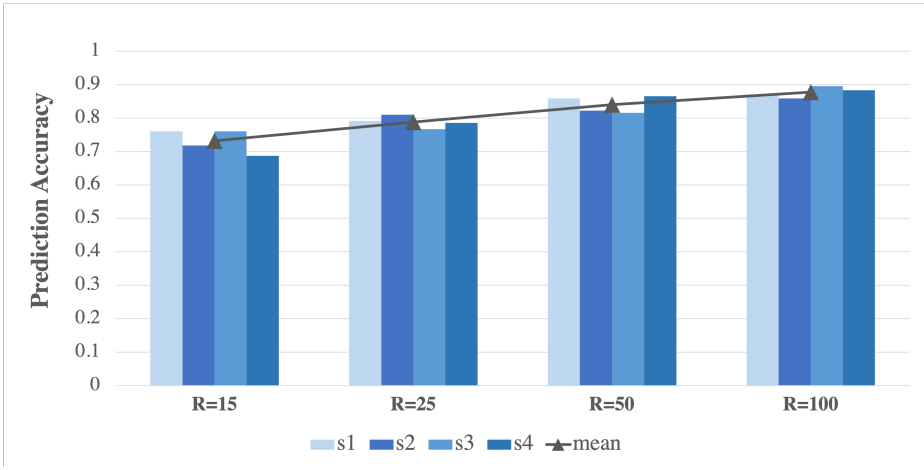


Figure 9: Prediction accuracy of sex using SVM with regions from 15 to 100. The dark grey triangle stands for the mean prediction accuracy over the four sessions, which also increased from 73% to 88% with the increase of regions.

4 Discussion

In terms of reliability, local and global reliability have different trends with the change in the resolution of the ICA-based pacellation. For ICC, it is hard to distinguish networks that have high reliability with those have poor reliability when there are only 15 regions. For 50 and 100 regions, we are able to identify networks with relatively high ICC, indicating their connectivity is quite consistent. High local reliability was observed within the visual network, cerebellum network, frontoparietal network and DMN. In addition, the connectivity between cerebellum and other networks has a relatively high reliability, such as frontoparietal and DMN. However, it becomes harder to distinguish when there are more than 200 regions, which may indicate that image is divided in too many regions than needed. As a result, moderate regions may be an optimal choice so that it is able to identify activated networks better. On the other hand, the increase in resolution reduces I2C2 significantly. The reason for that differences may be more regions provide more information of functional connectivity in brain and help identify more reliable connectivity between networks, while more regions also induce more variability, which leads to a decrease in global reliability. So we need to differentiate local and global reliability when we want to measure that of images.

Higher global reliability was observed on sessions on the same day than those on different days. For I2C2, sessions on the same day (i.e., session 1 & session 2) yield relatively higher I2C2 than sessions on different days (i.e., session 1 & session 3). In addition, higher local and global reliability were observed Comparing **Fig.3** with **Fig.A.2** and **Fig.A.3**, two sessions on the same day have higher ICC than all four sessions. Similarly, higher I2C2 was observed on two sessions on the same day than all four sessions. On the other hand, KS tests shown that connectivity matrices are

more similar between different sessions than between different subjects, which makes fingerprinting using connectivity matrix reasonable.

For functional connectivity fingerprinting, accuracy increases with the resolution, though 200 regions and 300 regions have pretty close performance. The method that combine two comparisons over two days as the identification rules yields the highest identification accuracy of over 90% with 300 regions. The method using two sessions on one day as the target set and sessions on the other day as the database set also gives relatively high identification accuracy. As a result, including more sessions in fingerprinting appears to improve the accuracy. In addition, the accuracy of methods using sessions on the same day is higher than the ones using sessions on different days, which is consistent with the global reliability measurement using I2C2. Switching the testing and database session only lead to small changes on identification performance.

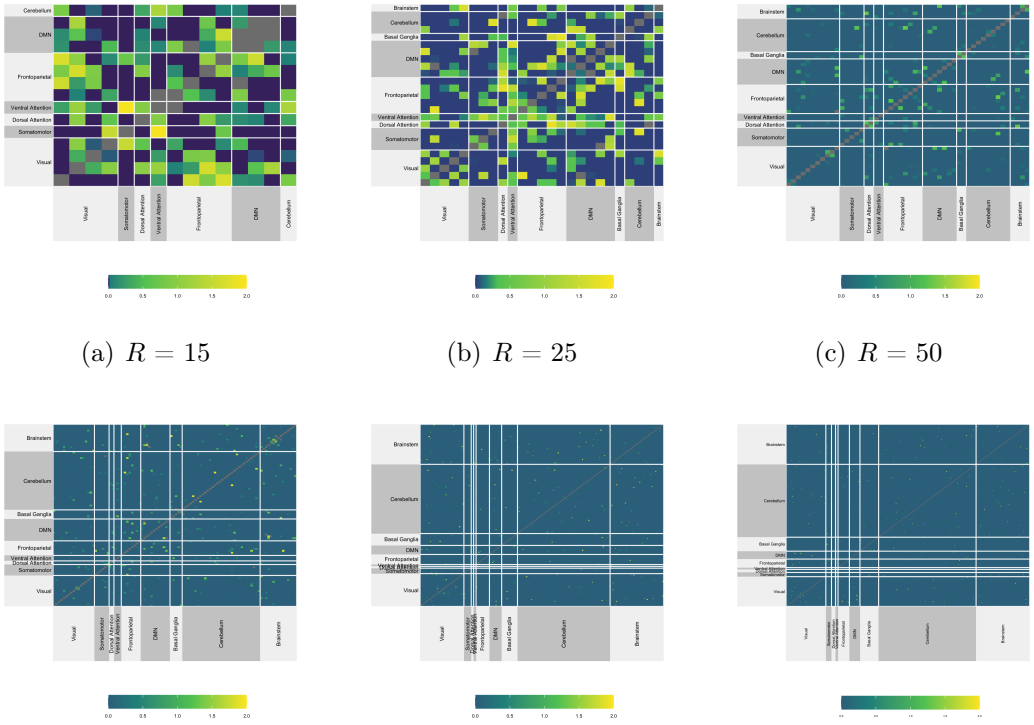
We were also interested in determining how useful the connectivity matrix was in predicting an individual's sex. Prediction performance is quite consistent among the four sessions using penalized models and SVM, and performance continuously improved to over 90% with the increase in resolution. However, after 100 regions there is only little improvement in prediction accuracy using penalized models, which may indicate 100 regions are able to provide the majority of features for prediction of sex. On the other hand, mean ICC decreases greatly with the increase of regions, regardless of whether the feature has a significant β or not in the prediction model. Since there are no obvious differences of ICC between features with significant β and non significant β, we could infer that more reliable features are not necessarily significant in the prediction of sex.

In summary, individual patterns of functional connectivity do exist. First, KS test

shown greater similarity between different sessions than different subjects. Second, fingerprinting and prediction sex using connectivity matrix yield high prediction accuracy. In addition, we also found some evidence that the resolution of the parcellation matters. According to the heat map of ICC and prediction performance of sex using penalized models, using a moderate number of regions seem to be an optimal choice, such as 100 regions. It provides more information and significant features of functional connectivity, which improves the performance of prediction. In addition, it avoids unnecessary computational cost, which only contribute little improvement to prediction accuracy.

A Appendix

Fig.A.1 displayed heat map of the absolute value of β in lasso models at different resolutions. With the increase in resolution from 15 to 300, the significant points on the heat map become increasingly sparse. Compared with the heat maps of ICC, the significant β distribute uniformly in networks instead of concentrating in several specific networks, while heat maps of ICC have some networks with higher values than others.

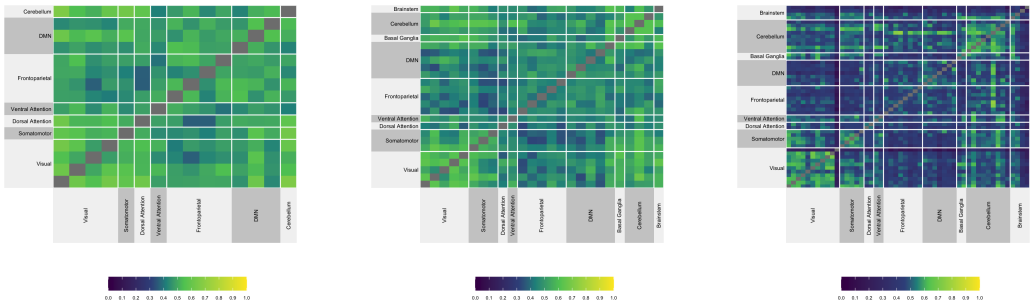


(a) $R = 15$ (b) $R = 25$ (c) $R = 50$

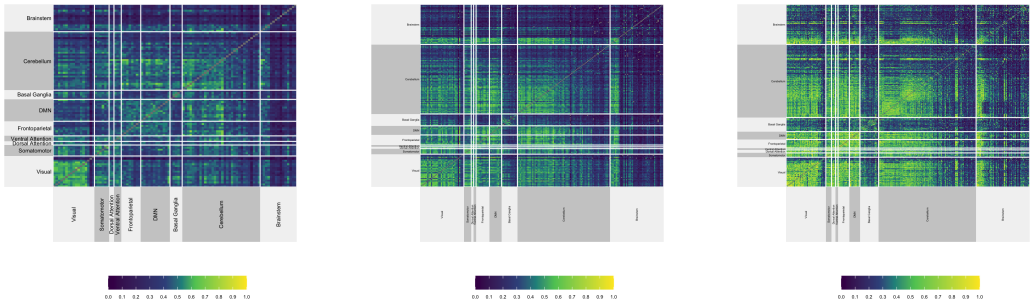
(d) $R = 100$ (e) $R = 200$ (f) $R = 300$

Figure A.1: Heat map of absolute value of β in lasso models with different regions. With the increase in resolution from 15 to 300, the significant points on the heat map become increasingly sparse. Compared with the heat maps of ICC, significant β distribute uniformly in networks instead of concentrating in several specific networks, while heat maps of ICC have some networks with higher values than others.

Fig.A.2 displayed heat map of ICC over session 1 & session 2 with different resolutions, while **Fig.A.3** displayed heat map of ICC over session 3 & session 4. They are similar with **Fig.3**, which is the heat map of ICC over all four sessions.

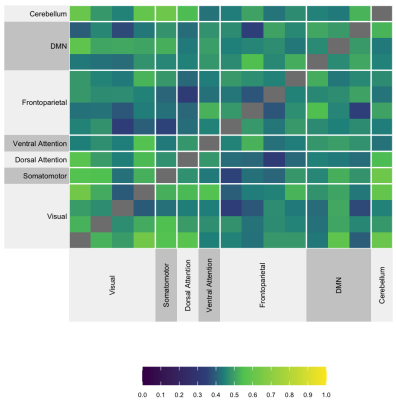


(a) $R = 15$ (b) $R = 25$ (c) $R = 50$

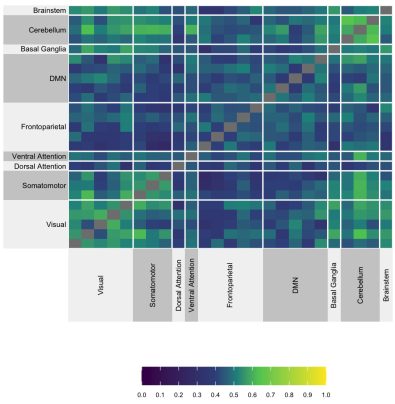


(d) $R = 100$ (e) $R = 200$ (f) $R = 300$

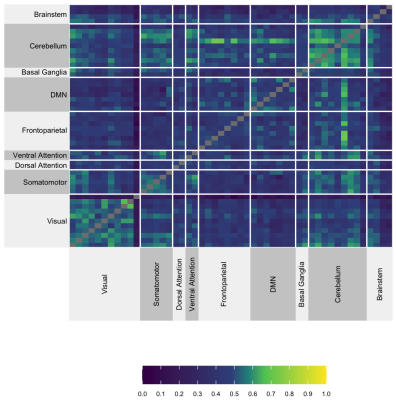
Figure A.3: Heat map of ICC comparing session 3 & session 4 different regions. Blue stands for poor reliability of connectivity, while yellow stands for high reliability. When we have less regions, it is more homogeneous and is hard to distinguish the networks with high reliability between the ones with poor reliability. With more regions, we are able to identify the networks with relatively high reliability. High ICC was observed within vision, cerebellum, frontoparietal and DMN.



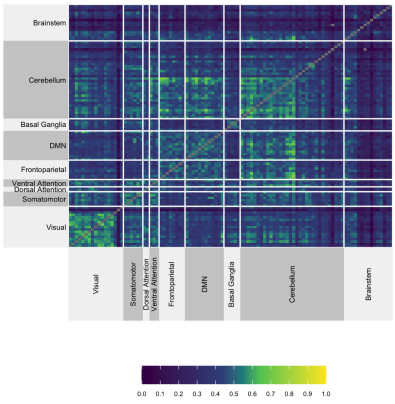
(a) $R = 15$



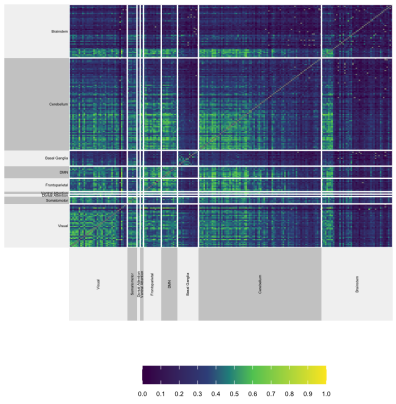
(b) $R = 25$



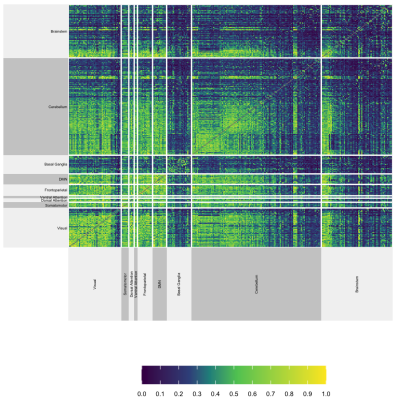
(c) $R = 50$



(d) $R = 100$



(e) $R = 200$



(f) $R = 300$

Figure A.2: Heat map of ICC comparing session 1 & session 2 at different resolutions. Blue stands for poor reliability of connectivity, while yellow stands for high reliability. When we have less regions, it is more homogeneous and is hard to distinguish the networks with high reliability between the ones with poor reliability. With more regions, we are able to identify the networks with relatively high reliability. High ICC was observed within vision, cerebellum, frontoparietal and DMN.

References

[1] Alexandre Abraham et al. "Extracting brain regions from rest fMRI with total-variation constrained dictionary learning". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2013, pp. 607–615.

[2] Christian F Beckmann and Stephen M Smith. "Probabilistic independent component analysis for functional magnetic resonance imaging". In: *IEEE transactions on medical imaging* 23.2 (2004), pp. 137–152.

[3] Craig M Bennett and Michael B Miller. "How reliable are the results from functional magnetic resonance imaging?" In: *Annals of the New York Academy of Sciences* 1191.1 (2010), pp. 133–155.

[4] Thomas Blumensath, Timothy EJ Behrens, and Stephen M Smith. "Resting-state FMRI single subject cortical parcellation based on region growing". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 188–195.

[5] Jason W Bohland et al. "The brain atlas concordance problem: quantitative comparison of anatomical parcellations". In: *PloS one* 4.9 (2009), e7200.

[6] Raymond J Carroll et al. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

[7] R Cameron Craddock et al. "A whole brain fMRI atlas generated via spatially constrained spectral clustering". In: *Human brain mapping* 33.8 (2012), pp. 1914–1928.

[8] Diane E Duffy and Thomas J Santner. "On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models". In: *Communications in Statistics-Theory and Methods* 18.3 (1989), pp. 959–980.

[9] Keith J Duncan et al. “Consistency and variability in functional localisers”. In: *Neuroimage* 46.4 (2009), pp. 1018–1026.

[10] Maxwell L Elliott et al. “What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis”. In: *Psychological Science* 31.7 (2020), pp. 792–806.

[11] Massimo Filippi et al. “The organization of intrinsic brain activity differs between genders: A resting-state fMRI study in a large cohort of young healthy subjects”. In: *Human brain mapping* 34.6 (2013), pp. 1330–1343.

[12] Nicola Filippini et al. “Distinct patterns of brain activity in young carriers of the APOE-ε4 allele”. In: *Proceedings of the National Academy of Sciences* 106.17 (2009), pp. 7209–7214.

[13] Emily S Finn et al. “Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity”. In: *Nature neuroscience* 18.11 (2015), pp. 1664–1671.

[14] WA Fuller. “Measurement Error Models, New York: JohnWiley”. In: *FullerMeasurement Error Models1987* (1987).

[15] Matthew F Glasser et al. “The minimal preprocessing pipelines for the Human Connectome Project”. In: *Neuroimage* 80 (2013), pp. 105–124.

[16] Roland H Grabner et al. “Individual differences in mathematical competence predict parietal brain activation during mental calculation”. In: *Neuroimage* 38.2 (2007), pp. 346–356.

[17] Ludovica Griffanti et al. “ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging”. In: *Neuroimage* 95 (2014), pp. 232–247.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[19] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[20] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[21] Vesa Kiviniemi et al. "Functional segmentation of the brain cortex using high model order group PICA". In: *Human brain mapping* 30.12 (2009), pp. 3865–3886.

[22] Terry K Koo and Mae Y Li. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research". In: *Journal of chiropractic medicine* 15.2 (2016), pp. 155–163.

[23] Saskia Le Cessie and Johannes C Van Houwelingen. "Ridge estimators in logistic regression". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41.1 (1992), pp. 191–201.

[24] John Mazziotta et al. "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1412 (2001), pp. 1293–1322.

[25] Kenneth O McGraw and Seok P Wong. "Forming inferences about some intraclass correlation coefficients." In: *Psychological methods* 1.1 (1996), p. 30.

[26] NeuroData et al. "Optimal Design for Discovery Science: Applications in Neuroimaging". In: (Aug. 2015). DOI: 10.6084/m9.figshare.1515021.v1. URL: https://figshare.com/articles/poster/Optimal_Design_for_Discovery_Science_Applications_in_Neuroimaging/1515021.

[27] Sharlene D Newman et al. "Frontal and parietal participation in problem solving in the Tower of London: fMRI and computational modeling of planning and high-level perception". In: *Neuropsychologia* 41.12 (2003), pp. 1668–1682.

[28] Jose Manuel Pereira, Mario Basto, and Amelia Ferreira da Silva. "The logistic lasso and ridge regression in predicting corporate failure". In: *Procedia Economics and Finance* 39 (2016), pp. 634–641.

[29] Stuart J Ritchie et al. "Sex differences in the adult human brain: evidence from 5216 UK biobank participants". In: *Cerebral Cortex* 28.8 (2018), pp. 2959–2975.

[30] SA Rombouts et al. "Test-retest analysis with functional MR of the activated area in the human visual cortex." In: *American journal of neuroradiology* 18.7 (1997), pp. 1317–1322.

[31] Gholamreza Salimi-Khorshidi et al. "Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers". In: *Neuroimage* 90 (2014), pp. 449–468.

[32] RL Schaefer, LD Roi, and RA Wolfe. "A ridge logistic estimator". In: *Communications in Statistics-Theory and Methods* 13.1 (1984), pp. 99–113.

[33] Lubdha M Shah et al. "Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state". In: *Brain and behavior* 6.5 (2016), e00456.

[34] David W Shattuck et al. "Construction of a 3D probabilistic atlas of human cortical structures". In: *Neuroimage* 39.3 (2008), pp. 1064–1080.

[35] H Shou et al. "Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2)". In: *Cognitive, Affective, & Behavioral Neuroscience* 13.4 (2013), pp. 714–724.

[36] Noah Simon et al. “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent”. In: *Journal of Statistical Software* 39.5 (2011), pp. 1–13. URL: https://www.jstatsoft.org/v39/i05/.

[37] Stephen M Smith et al. “Resting-state fMRI in the human connectome project”. In: *Neuroimage* 80 (2013), pp. 144–168.

[38] Bertrand Thirion et al. “Which fMRI clustering gives good brain parcellations?” In: *Frontiers in neuroscience* 8 (2014), p. 167.

[39] Yingjie Tian, Yong Shi, and Xiaohui Liu. “Recent advances on support vector machines research”. In: *Technological and economic development of Economy* 18.1 (2012), pp. 5–33.

[40] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[41] Nathalie Tzourio-Mazoyer et al. “Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain”. In: *Neuroimage* 15.1 (2002), pp. 273–289.

[42] Gaël Varoquaux et al. “A group model for stable multi-subject ICA on fMRI datasets”. In: *Neuroimage* 51.1 (2010), pp. 288–299.

[43] BT Thomas Yeo et al. “The organization of the human cerebral cortex estimated by intrinsic functional connectivity”. In: *Journal of neurophysiology* (2011).